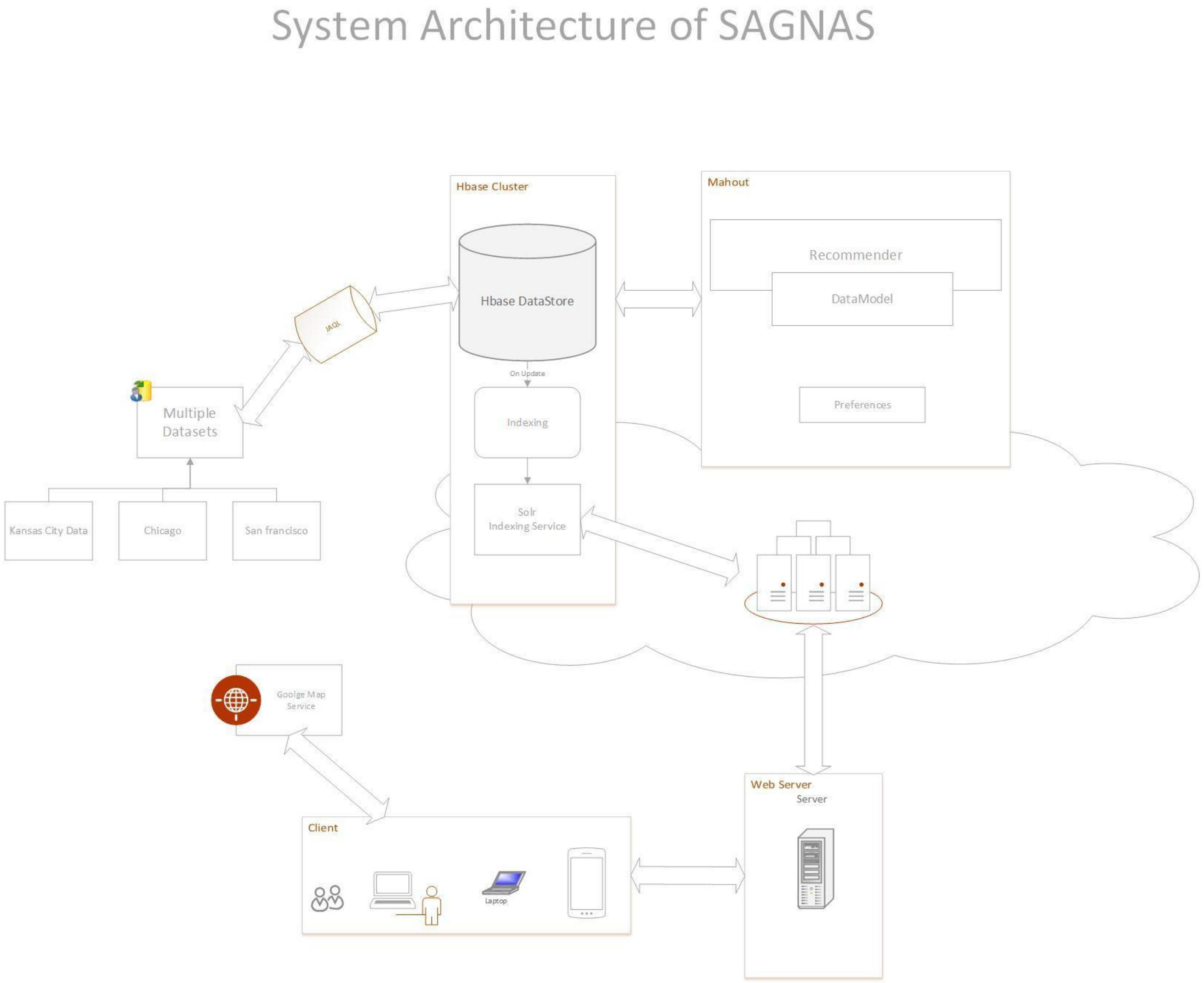
3-11-2014

SAGNAS (Group 9) First Increment

1) Framework Specification:



- **Domain Model**

  Data Sources

  We will need to store two different types of datasets:

  - The crime event data set, which can be used to display specific crime events as well as general crime statistics. This dataset is taken from city government websites, and is used to populate our analytical, Mahout generated data.

- The Mahout generated dataset, which will use various attributes within the crime data set to generate various analytical information that can help the user be safer.

Methodologies and Algorithms

We will use clustering to group zipcode-related crimes by time (either AM or PM) and by data (monthly).

Bayesian classification can be used to determine, given a specified date, time and crime type, whether a crime is more or less likely to happen.

Analytic Tools

Our main tool to analyze our data will be Mahout and its machine learning algorithms library.

Analytical Tasks

All of our analysis will be done on our crime datasets. Some of Mahout's algorithms can be run either periodically, such as the clustering algorithms, while creating Bayesian classifications will need to be done in real time based off given a data point.

Application Specification

Features List

New Services to be built:

- User recommendation service
  - Recommendations will be calculated weekly (not real-time)
  - Calculations will take the crime event dataset as an input, and produce a user recommendation dataset as an output
  - Will notify users of rate of crime happening in a certain area
  - Will utilize Mahout generated recommendation dataset accessed from HBase via Solr

- Crime information service
  - The input it uses is the crime event dataset
  - Will be displayed using charts and tables, using information based on the crime event dataset stored in HBase, indexed in Solr

- Crime display service
  - Will be displayed on Google Maps, using information based on the crime event dataset stored in HBase, indexed in Solr

- Ideally crime events would be displayed by zipcode, but due to various difficulties our plan instead is to display all events

Design of Mobile Client

We will utilize HTML5, Twitter Bootstrap, AJAX and CSS3

Implementation

Implementation of data model and algorithms

Our data model stores data produced from machine learning into Solr for easy client access. The crime data is stored on our server, with a structure that allows for straightforward Mahout usage and simple Solr document creation with the analytical findings. The crime data will interact with multiple machine learning algorithms via Mahout, including Bayesian classification and clustering.

We've faced challenges with implementation of our architecture due to issues with the integration of HBase-Mahout-Solr, and resolving this problem is our top priority going forward.

Implementation of services

We are going to use in-built services provided by Solr and HBase for data retrieval.

1. Solr services are used for faster data retrieval so all the analytical information is stored in Solr documents.
2. A Service is used to initiate the indexing process of HBase data and store the values in HBase.

Implementation of user interface

Our user interface is designed for mobile devices. We are utilizing twitter bootstrap to maintain a clean interface, and JSP to ultimately display the client-side code we have written.

HBase Schema Structure

```
{
   key:string,
   cf1: {
            address:string,
            crime:string,
            longitude:string,
            latitude:string
         },
   cf2: {
            date:string,
            id:string,
            time:string,
            zipcode:string
         }
}
```

<u>**Solr Doc Representation**</u>

```
Zip Code {
        Safety Parameter,
        Duration
        Type of Crime {
                type,
                count
                Report Numbers
        }
}
```

Data Model Clustering Mechanism:
Clustering Mechanism is based on the safety of a specific region which is decided based on the number of crimes took placed over the past few years.

The safety parameter is grouped as:

- Relax - Number of Crimes < 20
- Safe - 20 < Number of Crimes < 50
- Moderate - 50 < Number of Crimes < 100
- Less Dangerous - 100 < Number of Crimes < 200
- Dangerous - Number of Crimes > 200

The decision is made based on the Number of crimes took place in the region and the time period like a specific month or year.

Prediction Parameters:

Input Parameters: Zip Code(Default Current Location), Crime Type, Duration.

Output Parameters: Based on Crime rate our system is going to help the user to predict the Occurance of Crime during a certain period.

Project Management

List of tasks is listed with the status of what is being completed, under review, doing and To-do. All the issues are described in detail in scrumDo and the link to the account is http://www.scrumdo.com/projects/project/kdm-sagnas/summary